THE USE OF PROSODY IN SYNTACTIC DISAMBIGUATION

Patti Price (SRI), Mari Ostendorf (BU), Stefanie Shattuck-Hufnagel (MIT), Cynthia Fong (BU)

SRI International, Menlo Park, CA 94025
Boston University, Boston, MA 02215
MIT, Cambridge, MA 02139

ABSTRACT

Prosodic structure and syntactic structure are not identical; neither are they unrelated. Knowing when and how the two correspond could yield better quality speech synthesis, could aid in the disambiguation of competing syntactic hypotheses in speech understanding, and could lead to a more comprehensive view of human speech processing. In a set of experiments involving 35 pairs of phonetically similar sentences representing seven types of structural contrasts, the perceptual evidence shows that some, but not all, of the pairs can be disambiguated on the basis of prosodic differences. The phonological evidence relates the disambiguation primarily to boundary phenomena, although prominences sometimes play a role. Finally, phonetic analyses describing the attributes of these phonological markers indicate the importance of both absolute and relative measures.

INTRODUCTION

The syntax of spoken utterances is frequently ambiguous. Yet listeners usually arrive at something close to the intended meaning. Information listeners might use in disambiguation includes knowledge of the world, shared context, and a source of non-syntactic information that is under-represented in written communication: the prosody of the utterance. By 'prosody' we mean suprasegmental information in speech, such as phrasing and stress, which can alter perceived sentence meaning without changing the segmental identity of the components.

Since prosody plays an important role in speech communication, a clear understanding of the mapping between prosodic and syntactic structure would reveal significant aspects of the cognitive processes of speech production and perception. In addition, it would provide guidelines for the synthesis of more natural-sounding speech. Further, any contribution that prosody can make to the resolution of structural ambiguities will be particularly helpful in spoken-language understanding, where lexical and structural ambiguities of written forms are compounded by difficulties in finding word boundaries and in identifying words reliably in automatic speech recognition. Here, we study the mapping between prosody and syntax by minimizing the contribution of other possible cues to the resolution of ambiguity. This study forms the foundation for further work on modeling prosody by assessing a set of syntactic environments in which prosody alone might be used to disambiguate sentences, and by analyzing the correspondence between the phonological and phonetic attributes of the prosodic structure of utterances and their perceived meanings.

We begin by discussing previous work on the relationship between prosody and syntax. We then describe the recording of the corpus, and present results for the experimental studies which consider: (1) the accuracy and confidence of listeners in disambiguating different types of syntactic structures, (2) the phonological analysis of prosodic cues associated with the different structures, and their relation to the disambiguation results, and (3) a phonetic analysis of the phonological markers. Finally, we discuss the implications of these results, and raise some unresolved questions that suggest directions for future research.

BACKGROUND

With few exceptions (e.g., [9]), previous studies have focussed either on relating phonological aspects of prosody to syntax (e.g., [8], [14],[3], [12]), or on relating phonetic/acoustic evidence to syntax and perceived differences (e.g., [19], [4], [20], [7], [11], [6], [21]). A few studies, e.g., [16], have considered the mapping from phonology to acoustics. The more phonetic/acoustic studies typically used a small number of minimal pairs of utterances in order to facilitate the acoustic measurements and to control parameters more precisely (exceptions include [10], and [5] where larger data sets were used). In contrast, the more phonological studies have focussed either on 'illustrative examples' or on text to which prosodic markers have been assigned on the basis of the syntax of the sentence. These studies have typically ignored the fact that there are several possible prosodic choices for a given syntactic structure. The focus in recent theoretical linguistics on human *competence* for language production, has resulted in neglect of actual language production and neglect of an area required for speech understanding (by human or by machine): the mapping from acoustics to meaning. Clearly, speech communication involves both production and perception, and it involves performance as well as competence.

The work presented in this paper extends previous work, including the important contribution of [13], in several ways. First, focussing only on surface-structure ambiguities (since earlier work indicates that these are good candidates for disambiguation), we investigate the ability of listeners to disambiguate sentences for different types of syntactic structures, using several instances of each type. Second, our focus here is on both production and perception. We tried to avoid exaggeration of any disambiguating strategies on the part of speakers and listeners by separating the ambiguous pairs from each other in time (no two members of an ambiguous pair occurred in the same session either for speakers or for listeners). Third, to increase reliability without assessing a large pool of subjects, we used four professional FM radio announcers, who have proved to be very consis-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1991		**00-00-1991 to 00-00-1991**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Use of Prosody in Syntactic Disambiguation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Computer Science,Spoken Language Systems Group,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**6**	

tent speakers in our pilot studies. Fourth, in analyzing the cues used in disambiguation, we have investigated the possible use of prominence associated with pitch accents, in addition to prosodic phrase boundary cues. Finally, to compare durational structures across the various sentences used, and to facilitate generalization beyond the specific sentences used, we present results in terms of relative, rather than absolute, durational patterns. By combining phonological analyses of prosodic elements such as boundary tones and prominences with investigation of their acoustic correlates and their perceptual effects, we hope to shed some light on both the mapping between syntactic and prosodic structure, and on the role of prosody in resolving various types of syntactic ambiguity.

CORPUS

Our methodology involved (1) recording pairs of structurally ambiguous sentences, (2) presenting the resulting utterances to naive listeners for perceptual judgements, and (3) comparing the phonological and phonetic characteristics of the spoken utterances with listeners' ability to disambiguate them. The recordings, which formed the basis for both perceptual experiments and phonetic and phonological analyses, are described below.

We used 35 sentences pairs, ambiguous in that the two members of each pair contained the same string of phones, and could be associated with two contrasting syntactic bracketings. The sentences manifested seven types of structural ambiguity:

(1) parenthetical clauses vs. non-parenthetical subordinate clauses,

(2) appositions vs. attached noun (or prepositional) phrases,

(3) main clauses linked by coordinating conjunctions vs. a main clause and a subordinate clause,

(4) tag questions vs. attached noun phrases,

(5) far vs. near attachment of final phrase,

(6) left vs. right attachment of middle phrase, and

(7) particles vs. prepositions.

Note that "high vs. low" attachment is probably a more accurate syntactic description than "far vs. near" attachment. However high vs. low attachment could involve the same site in the string of words being parsed, and our instances of far (high) attachment all involve attachment to phrases ending in a word that is not neighboring the word to be attached. Therefore, we instead use the more descriptive terms "far" and "near".

In each of the 7 categories, there were 5 pairs of ambiguous sentences. In presentation, each sentence was preceded by a disambiguating context of one or two sentences. The target sentences were fully voiced to facilitate pitch tracking for acoustic analysis. We use the term *size of syntactic break* to reflect the number of syntactic brackets that would occur between two pairs of words: more brackets correspond to a larger syntactic break. The site with the largest number of brackets is *the major syntactic break.* For structural categories 1-4, sentence A of the pair involved a larger syntactic break than sentence B. For the attachment ambiguities 5-7, sentence A of the pair had the larger syntactic break later in the sentence than did sentence B.

The sentences were recorded by four professional FM public radio newscasters, one male and three female, who were naive with respect to the purposes of the experiment. The newscasters were asked to read the sentences in context, using their standard radio style of speaking. In a pilot study, we found the FM radio style to have more clearly and consistently marked prosodic cues than a non-professional speaking style [18]. Our hope was that this style would be easier to label prosodically, and therefore the contributions of specific phonological cues would be easier to identify.

The announcers were presented with the written sentences in context paragraphs, with the sentence types and A/B members of the pairs assigned to two recording sessions, so that the two contrasting members of a pair did not occur in the same session. The speakers were not told that there were special target sentences within the paragraphs. The recording sessions were separated by at least a few days and often several weeks, to minimize the possibility that the announcers would produce unnatural versions in an attempt to emphasize potential differences between the two members of a pair.

Our goal was to create sentence pairs that were segmentally identical but syntactically different, so that we could investigate the relationship between syntax and prosody independent of any differences contributed by the segments. Although they were not prosodically incorrect, tag sentences in which the tags were read as questions were rerecorded as statements so that the question boundary tone cue would not confound the potential contribution of other prosodic cues.

PERCEPTUAL EXPERIMENTS

Methods

For the perceptual experiments, the spoken context sentences were edited out so that the target sentences could be presented in isolation. The 35 sentence pairs produced by a single speaker were presented to listeners in two sessions; only one member of each pair was heard in each session using a mixed assignment of half type A and half type B sentences in each session (analogous to the strategy used for recording the sentences). The different syntactic types were interleaved, and A versions always appeared before B versions on the answer sheet. The listeners heard the sentences in a small conference room from a portable stereo. The tape player was stopped between sentences until subjects were ready to continue; the subjects were under no time constraints to make their judgements. Each listening session (35 sentences) took approximately 40 minutes, and was conducted without any additional breaks. Listening sessions were separated by at least three weeks to minimize listener recall of the previous session's sentences. Listeners were given an answer sheet with both disambiguating contexts written out for each sentence; the target sentence was printed in bold at the end of each context. They were asked to mark the context which they thought best matched what they heard, with an additional marker if they were confident of their decision. Subjects were rewarded with pizza and soft drinks after the session.

The subjects were all native speakers of American English, naive with respect to the purpose of the experiments. Most were engineering students, recruited through flyers advertising the free pizza. For the second two speakers, to attract more subjects, we increased the incentive by offering an additional $50 prize to the person who scored highest on this task. The number of listeners who heard both sessions for each of the different speakers was 13 for Speaker F1A, 15 for F2B, 17 for F3A and 12 for M1B. Different subjects partici-

pated in the experiments for the different speakers, although there was some overlap in the subject pool. Four subjects participated in all four experiments.

Results

For the analysis, we assume that the speaker produced the intended version of the sentence, and define a correct listener response as one which identifies that version. Accuracy is the percentage of correct listener responses. Confidence is the percent of the time that listeners indicated that they were confident of the response choice. Table 1 summarizes average subject accuracy for the different types of ambiguity. The averages are taken over the four speaker averages, so as not to more heavily weight the utterances that were heard by more listeners. The averages for each speaker are taken across five versions of each structural type, as well as across the various listeners (12-17 per talker).

Table 1 shows that subjects could reliably disambiguate many, but not all of the ambiguities. Subjects were rarely confident *and* incorrect, and the confidence is somewhat correlated (0.64) with the accuracy. On the average, subjects did well above chance (84% correct) in assigning the sentences to their appropriate contexts, although subjects were confident of their judgments only 52% of the time. Also on average, main-subordinate (3B) sentences and near attachments (5B) were close to the chance level; parentheticals (1A), far attachments (5A) and non-tags (4B) were recognized at levels greater than chance but not reliably; and all other sentence types were reliably disambiguated.

Type	Version A	Version B	Overall
1. Parenthetical or not	77	96*	86
2. Apposition or not	92*	91*	92
3. M-M vs. M-S	88*	54	71
4. Tags or not	95*	81	88
5. Far/near attachment	78	63	71
6. Left/right attachment	94*	95*	95
7. Particle/Preposition	82*	81*	82
Average	87	80	84

Table 1. Perceptual experiment results, averaged over the four speakers, for ambiguous sentence interpretation. The Version A/B figures are based on 285 total observations of each class. An asterisk marks the A and B version responses that had high accuracy in listener responses. (High accuracy was defined to be average accuracy minus the standard deviation greater than 50%.)

PHONOLOGICAL ANALYSIS

The perceptual experiments described above clearly show that speakers can encode prosodic cues to structural ambiguities in ways that listeners can use reliably. This section attempts to find a phonological answer to the question: How do they do it? To approach this question, we labeled discrete, prosodic phenomena (specifically, prosodic phrase boundaries and prominences) that could mark structural contrasts phonologically. We then analyzed the relationship between these labels and the patterns in the perceptual accuracy study. There are other prosodic cues (e.g., the *type* of pitch accent), and there are other phonological correlates of the prosodic structure (e.g., phonological processes at prosodic boundaries) which can likely play a role in disambiguation. However, analysis of these phenomena was beyond the scope of the present study. In the following section, we describe our labeling system and analyze the associated constituents in terms of their relationship to the syntactic structures in our corpus, and the accuracy with which sentences are identified.

Perceptual Labels

We chose labels based on three criteria: (1) they should be used consistently within and across labelers, (2) they should be rather close to surface forms (to make eventual automatic detection more tractable and to improve labeler consistency), and (3) they should provide a mechanism for communicating information to a parser. For these reasons, our notation differs somewhat from that of other systems, although it is similar in many respects.

We used seven levels to represent perceptual groupings (or, viewed another way, degrees of separation) between words. These seven levels appeared adequate for our corpus and also reflected the levels of prosodic constituents described in the literature. Our labeling experience led us to adopt the maximum number of levels suggested in the literature, although not all are universally accepted. We used numbers to express the degree of decoupling between each pair of words as follows: 0 - boundary within a clitic group, 1 - normal word boundary, 2 - boundary marking a grouping of words generally having only one prominence, 3 - intermediate phrase boundary, 4 - intonational phrase boundary, 5 - boundary marking a grouping of intonational phrases, and 6 - sentence boundary.

Break indices of 4, 5, and 6 are "major" prosodic boundaries; constituents defined by these boundaries are often referred to as 'intonation phrases' (e.g., see [2]), and are marked by a boundary tone. Boundary tones were labeled using two types of falls (final fall and non-final fall), and two types of rises (continuation rise and question rise). The break index 3 corresponds to the unit referred to as an 'intermediate phrase' in [2] or a 'phonological phrase' in [14]. The 'phrase accent' pitch marker theoretically associated with the intermediate phrase was not labeled.

Prominent syllables in the sentences were labeled using P1 for major phrasal prominence; P0 for a lesser prominence; and C for contrastive stress, which occurred rarely in these sentences (marked on 1% of the total words for four speakers).

The prosodic cues were labeled perceptually by three listeners using multiple passes. The data were first labeled by the listeners individually; any differences in markings were then discussed; and then the sentence was replayed a few times to allow the labelers to revise their markings. Finally, a majority vote of the labels (which at this point had a correlation of 0.96 across labelers) was used as the final hand-marked label set. All labeling was perceptual.

Analysis

To separate semantic effects from effects that should occur throughout the syntactic class, we paid particular attention to those cues that

reliably occurred in the A versions of one class, but never in the contrasting B versions, or vice versa. We also paid particular attention to those sentences that had high accuracy and confidence and to the outlier sentences. Below we mention some general results and then discuss briefly the individual classes investigated.

General Observations: We found that prosodic boundary cues are associated with almost all reliably identified sentences. Presence of an intonational phrase boundary (break index 4 or 5) was often, but not always, a reliable cue and was most often observed at embedded or conjoined clause boundaries (marked by commas in the text). In addition, a difference in the relative size of prosodic break indices, or in the location of the largest break regardless of size, was frequently the only disambiguating information in the labels for the smaller syntactic constituents that were reliably disambiguated. By and large, relatively larger break indices tended to mean that syntactic attachment was higher rather than lower. In contrast to the pervasive association of boundary cues with successful disambiguation, prominence seemed to play mainly a supporting role, and was the sole cue in only a few sentences.

Parenthetical (A) vs. non-parentheticals (B): The A versions always have break indices larger than 3 surrounding the parenthetical, except for one talker's rendition of one sentence. The B members have break indices less than 4 at one or both of the corresponding sites. In all cases, the sentences with major prosodic breaks surrounding the parenthetical were identified as version A by 75% or more listeners, and sentences without the major prosodic breaks were identified as version B 80% of the time or more. This generalization includes an anomalous A version having a 3 at the parenthetical boundary, which was identified in accordance with the indices rather than in accordance with the speaker's intent.

Apposition (A) vs. non-apposition (B): The A version of the pair, the appositive, always has a major prosodic break both before and immediately following the appositive. The B version of the pair typically has a small break index at one or both of the corresponding sites. Two speakers produced a major break at the 'wrong' location, i.e., after "are" in "Wherever you are in Romania or Bulgaria, remember me." This predicts that the sets should be clearly separable, except for this sentence, which is what we found: All were labeled by the naive listeners at 87% accuracy or higher, except for this sentence, which was 73% correct.

Main-main (A) vs. main-subordinate sentences (B): The A versions of the pairs were typically well-identified, whereas the B versions tended to be close to the chance level. This could be the result of a syntactic response bias if the conjunction constructions are preferred over the deleted "that" in the alternants. This is interesting since the bracketings differ for the two versions of the sentence, and yet the two versions are apparently not well separated perceptually. The prosodic transcriptions suggest a reason: *both* versions of the sentence have a major prosodic boundary in the same location, associated with the embedded (B) or conjoined (A) sentence.

Tags (A) vs. non-tags (B): The A members all have a major prosodic break before the tag, and these were all identified as A versions (92% or more or the time). One talker produced one B version with a major prosodic boundary in the "wrong" place, and 92% of the listeners identified this utterance as version A, in accordance with the prosody. Two other B versions were frequently misidentified; these sentences had no boundary tone, but did have a break index of 3 (the largest in these sentences) at the site corresponding to the boundary of the tag.

Far (A) vs. near (B) attachment sentences: The A versions showed a tendency to have the largest break index in the sentence before the phrase to be attached to a "far" site (i.e., a site other than to a phrase ending in the immediately preceding word). This pattern occurred in 15 of the 20 A utterances and only one of the B utterances. One talker's production of one A version had a 2 at the site in question, and a majority of the listeners labeled this as version B, which happened with none of the other A versions. Thus, the location of a relatively large break index at the site in question appears to block the "near" (low) attachment, and a relatively small index appears to enhance it.

Left (A) vs. right (B) attachment sentences: For every rendition by every talker, there was a smaller break index at the attachment location than at the other end of the word or phrase to be attached. For the four sentence pairs that differed in comma location, the difference between the two break indices was large (2 or more), typically 0 or 1 in the location without a comma and 3, 4 or 5 in the location with the comma. These utterances were very reliably identified, with greater than 92% accuracy for all but one case.

Particles (A) vs. prepositions (B): There is less frequently a major prosodic break before a prepositional phrase compared to conjoined or embedded sentences: 60% of the prepositional phrases in this class followed a major prosodic break, compared to 90% observed in the context of clauses. The real structural clue appears to be not the absolute size of the break index but its relative size. For all A versions, we observed a smaller break index between the verb and particle, compared to the indices before the verb or after the particle. For the B versions, the relations were reversed: there was a tendency to have a larger break between the verb and preposition, compared to those before the verb or after the preposition.

There was little systematic difference in the speakers' use of prosodic cues. There were some differences in individual sentences which accounted for the variation in listener responses, but no consistent characteristics attributed to any one speaker. The correlation of break indices between pairs of speakers was 0.94-0.95, and the relative frequencies of prominences for the different speakers were also very similar. This result is consistent with the finding in [5] of a high correlation in duration patterns between different versions of the same utterance read by non-professional speakers.

PHONETIC ANALYSIS

We have thus far presented evidence that naive listeners can reliably use prosody to separate structurally ambiguous sentences, and phonological evidence that suggests how listeners might use prosody to assign syntactic structure. Other studies have focussed on syntactic differences associated with disambiguation. Our evidence shows that the prosodic structure can point to the syntactic differences in systematic ways: sentences with certain correspondences between syntactic and prosodic structures are reliably disambiguated, whereas others are not. In this section we investigate some of the phonetic evidence that might be responsible for the prosodic disambiguation. Since previous work suggests that the primary prosodic cues are duration and intonation, the present study is confined to these two cues. However, we acknowledge that other cues, such as the application or non-application of phonological rules, contribute to the perception of prosodic boundaries. We tried to minimize

such effects by asking the speakers to reread sentences in which overt segmental cues were produced, i.e., where the gross phonetic transcription of the two versions of the sentence would differ.

In the results presented here, segment duration normalization is determined automatically using an HMM-based speech recognition system, the SRI Decipher system, which uses phonological rules to generate bushy pronunciation networks that should enable more accurate phonetic transcription and alignment than single pronunciation speech recognizers [22]. Each phone duration was normalized according to speaker- and phone-dependent means as described in [15]. The variance of normalized duration in different contexts tends to be large, because the normalization has not accounted for effects such as syllable position, phonological and phonetic context, and speaking rate. In other work, we have found that variance can be reduced by adapting the phone means according to a local estimate of the speaking rate, which also plays a role in determining phoneme duration.

We observed longer normalized durations for phones preceding major phrase boundaries and for phones bearing major prominences compared to other contexts. As mentioned earlier, it has long been noted that syntactic breaks are often associated with duration lengthening in the phrase-final syllable, though the scope of the lengthening is in dispute. We measured average normalized duration in the rhyme of the final syllable of all words and found that higher break indices are generally associated with greater normalized duration. The fact that duration is affected by constituents at many levels in the prosodic hierarchy is interesting, and consistent with our observations that relative break index size is meaningful even below the level of the intonational phrase (4,5). However, more research is needed on this question, since only the difference between the groups 0-3 (without boundary tone) and 4-6 (with boundary tone) is statistically significant; differences within those groups are not. Pauses are also associated with major prosodic boundaries, occurring at 48/212 (23%) boundaries marked with 4 and 17/25 (67%) boundaries marked with 5. Sentence-final pauses could not be measured for these sentences, which were always the final sentence in a paragraph. In only one case did a pause occur after a 3.

Our analysis of normalized duration of the vowel nucleus for the different prominence markings revealed that: (1) major prominences (P1, C) tend to be longer than unmarked or minor (P0) prominences, although the effect is small before major prosodic breaks; (2) word-final syllables tend to be longer than non-word-final syllables; (3) syllables are longer in words before major breaks than before smaller breaks, though the effect is more dramatic for word-final syllables than for non-word-final syllables; and (4) the effects seem to be somewhat independent: the longest syllables are those with a major prominence, in word-final position, before a major break.

Intonational cues observed included boundary tones, pitch range changes and pitch accents. Boundary tones are involved for the break indices 4, 5 and 6. Sentence-final (6) boundary tones are typically final falls; level (5) boundary tones are usually perceived as incomplete falls; and intonational phrase (4) boundary tones are most often continuation rises but occasionally are perceived as partial falls. Tags were sometimes associated with a sentence-final question rise, though we tried to eliminate this cue as much as possible by asking the radio announcers to reread versions when this occurred. Another intonational cue was a perceived drop in pitch baseline and range in a parenthetical phrase, relative to the rest of the sentence. This pitch range change was not always perceived for appositives. In examining the associated fundamental frequency (F0) contours, we observed a region of reduced F0 excursion during the period of perceived range change. Though intonation is an important cue, duration and pauses alone provide enough information to automatically label break indices with a high correlation (greater than 0.86) to hand-labeled break indices [15].

Since prominence was not consistently associated with specific syntactic structures in any systematic pattern (with the exception of particles), it appears that the disambiguating role of prominences (or pitch accents) differs from that of boundary phenomena, being associated more with the semantics rather than with the syntax of an utterance. In other words, we suspect, with others, that prominence is related more to the contextual focus of the sentence.

DISCUSSION

We have confirmed that, for a variety of syntactic classes, but not all, naive listeners can reliably separate meanings on the basis of differences in prosodic information. We have further shown phonological and phonetic evidence bearing on how they might do this: by the tendency to associate relatively larger prosodic breaks with larger syntactic breaks. Further, syntactic boundaries of clauses that contain complete sentences nearly always coincide with the boundaries of major prosodic constituents (as marked, e.g., by syllable-final lengthening, a boundary tone and perhaps a pause). Syntactic constituents within these major constituents may be associated with any of several different levels of prosodic boundaries, i.e., speakers have more choice in phrasing, and prosodic boundaries need not correlate perfectly with syntactic ones, though they often do. We have also shown the importance of the *relative* size of prosodic breaks within a sentence. Though evidence relating to boundary phenomena appeared to be most important, there were some structures for which phrasal prominence either was the only cue or played a supporting role in distinguishing between the two versions.

Several aspects of the design of our experiment require comment involving the interpretation of our results. First, the disambiguation of some of the sentences may have been confounded by prosodic cues related to non-syntactic factors, e.g., given vs. new information, focus, contrastive stress, etc. However, the use of several sentences and of several speakers should minimize these effects, and should make it unlikely that there is a systematic correlation between such effects and the A and B versions of the sentences. Clearly, to fully elucidate the relationship between prosody and syntax will require the investigation of far more examples of far more syntactic constructions than we have been able to use in this study. Second, our finding of a correlation between the syntax and the phonological markers of prosody may have been corrupted by the fact that the labelers typically knew which version they were listening to. However, the labelers did not know the relative accuracy of the responses of the naive subjects. Therefore, these labels are relevant insofar as they account for both the accurate and the inaccurate responses. Third, we did not investigate the role of syntactic constituent length, which others have found to influence the placement of prosodic boundaries [1]. Lastly, the use of read speech by professional radio announcers as speakers raises questions about generalizing the results to spontaneous speech by more average talkers. We believe that the use of the professional speakers has

allowed us to obtain initial results using far fewer speakers than would be needed using non-professionals. We hypothesize that the prosodic cues will be similar for non-professional speakers, although less consistently used and not as clearly marked.

Our results have both theoretical and empirical implications. We have shown that naive listeners can use prosody to separate structurally ambiguous sentence pairs, and we have further shown phonological and acoustic evidence of how they might do this. In speech generation applications, such information is useful since different prosodic markers will affect the interpretation of a sentence. Prosodic cues are particularly important in computer speech understanding applications, where the semantic rules available to the system are limited relative to the capabilities of human listeners. In addition, in these applications, prosodic cues can be used prior to semantic analysis, to reduce the number of syntactically acceptable parses by eliminating those inconsistent with the prosody [15].

FUTURE DIRECTIONS

The results reported here provide evidence for some systematic relationships between prosody and syntax that should be explored further in several ways. First, a larger number of syntactic structures must be examined in order to make the prosody/syntax relationship more explicit. Second, we note that some sentences were successfully disambiguated with cues that were not represented in our labeling scheme. Since prominences were not differentiated as to type of pitch accent, a more detailed classification of intonation in such contexts could yield more information. Finally, for computer speech understanding applications, it will be important to investigate the extension of these results to spontaneous speech by non-professional speakers, where hesitation phenomena and speech errors will affect the prosodic structure.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IRI-8805680 in coordination with DARPA/NSF funding under NSF grant number IRI-8905249. The government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the government funding agencies. In addition, the authors wish to thank Andrea Levitt and Leah Larkey for their work with Patti Price in generating the ambiguous sentences; the radio announcers at WBUR in Boston who recorded the sentences; the subjects who participated in the perceptual experiments; Nanette Veilleux and Colin Wightman for many hours of prosodic labeling; and Gay Baldwin for verification of phonetic alignments. We thank John Bear for providing syntactic bracketings of the sentences and for many useful comments on the paper.

REFERENCES

[1] Bachenko, J. and Fitzpatrick, E. (1990). "A computational grammar of discourse-neutral prosodic phrasing in English," *Computational Linguistics* 16:3, 155-170.

[2] Beckman, M. and Pierrehumbert, J. (1986). "Intonational structure in Japanese and English," *Phonology Yearbook* 3, ed. J. Ohala, pp. 255-309.

[3] Bing, J. (1984). "A discourse domain identified by intonation," pp. 11-19 in *Intonation, Accent and Rhythm: Studies in Discourse Phonology* (New York, de Gruyter).

[4] Cooper, W. and Sorensen, J. (1977). "Fundamental frequency contours at syntactic boundaries," *J. Acoust. Soc. Am.* 62:3, 683-692.

[5] Crystal, T. H. and House, A. S. (1990). "Articulation rate and the duration of syllables and stress groups in connected speech," *J. Ac. Soc. Am.* 88:1, 101-112.

[6] Duez, D. (1985). "Perception of silent pauses in continuous speech," *Language and Speech* 28:4, 377-389.

[7] Garro, L. and Parker, F. (1982). "Some suprasegmental characteristics of relative clauses in English," *J. Phonetics* 10, 149-161.

[8] Gee, J. P. and Grosjean, F. (1983). "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology* 15, 411-458.

[9] Geers, A. (1978). "Intonation contour and syntactic structure as predictors of apparent segmentation," J. *Exp. Psych: Hum. Perc. and Perf.* 4:3, 273-283.

[10] Klatt, D. H. (1975). "Vowel lengthening is syntactically determined in a connected discourse," *J. Phonetics* 3, 129-140.

[11] Kutik, E., Cooper, W., and Boyce, S. (1983). "Declination of fundamental frequency in speakers' production of parenthetical and main clauses," *J. Ac. Soc. Am.* 73:5, 1731-1738.

[12] Ladd, D. R. (1986). "Intonational phrasing: the case for recursive prosodic structure," *Phonology Yearbook* 3, 311-340.

[13] Lehiste, I. (1973). "Phonetic disambiguation of syntactic ambiguity," *Glossa 7:2*, 107-121.

[14] Nespor, M. and Vogel, I. (1983) "Prosodic structure above the word," *Prosody: Models and Measurements*, Cutler and Ladd, eds. (Springer-Verlag), pp. 123-140.

[15] Ostendorf, M., Price, P., Bear, J. and Wightman, C. (1990). "The use of relative duration in syntactic disambiguation," *Proceedings of the 3rd DARPA Workshop on Speech and Natural Language.*

[16] Pierrehumbert, J. (1981). "Synthesizing intonation," J. *Ac. Soc. Am.* 70:4, 985-995.

[17] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (manuscript submitted) "The use of prosody in syntactic disambiguation."

[18] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Veilleux, N., (1988). "A Methodology for Analyzing Prosody," *J. Ac. Soc. Am.* 84, Suppl. 1, S99.

[19] Scholes, R. (1971). "On the spoken disambiguation of superficially ambiguous sentences," *Language and Speech* 14, 1-11.

[20] Thorsen, N. (1980). "A study of the perception of sentence intonation -- Evidence from Danish," *J. Ac. Soc. Am.* 67:3, 1014-1030.

[21] Thorsen, N. (1985). "Intonation and text in standard Danish," *J. Ac. Soc. Am.* 77:3, 1205-1216.

[22] Weintraub, M.*et al.* (1989), "Linguistic constraints in hidden Markov model based speech recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 699-702.